Meeting report

Biocomputing enters its adolescence

Shamil Sunyaev

Address: Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. E-mail: ssunyaev@rics.bwh.harvard.edu

Published: 31 May 2005

Genome ***Biology*** 2005, **6:**325 (doi:10.1186/gb-2005-6-6-325)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/6/325

A report on the tenth Pacific Symposium on Biocomputing, Big Island, Hawaii, USA, 4-8 January 2005.

This year's Pacific Symposium on Biocomputing saw a diverse group of computational biologists discussing an equally diverse collection of applications of computational methods to biology. At this tenth symposium under the Hawaiian sun, the young field of computational biology left its infancy behind and became a teenager. A unique feature of the Pacific Symposia is that session topics are selected from submitted proposals. This ensures that the conference is well tuned to the changing character of the field, and this year's symposium covered a very wide spectrum of biological problems of interest to those developing computational methods. Sessions on biogeometry - the application of computational genometry to three-dimensional structures of biopolymers - and the informatics of structural genomics reflect a long-standing interest of the Pacific Symposia, and of computationalists in general, in problems of structural biology. Other sessions focused on methods for combining heterogeneous data sources at a genome-wide scale, the use of biomedical ontologies to provide a structured and unified means of genome annotation, and genomic variation in populations and its implications for pharmacogenomics.

From structure to function

Sequence analysis remains a dominant method for predicting functional features of genes and proteins and for the annotation of genomes. As well as methods based on sequence similarity, other evolution-based methods relying on complete genome sequences are gaining ground. As David Eisenberg (University of California, Los Angeles, USA) noted in his keynote lecture, the power of computational methods for predicting protein interactions from genomic location and the coevolution of genes has been greatly increased as a result of the extraordinary growth of the number of complete genomes. This has allowed the development of new types of methods for detecting interactions based on the coevolution of triplets of genes rather than just of gene pairs.

Although it is obvious that the spatial structure of biomolecules contains much more information than the sequence, the practical use of structural data remains limited. An increasing number of proteins have a known structure but an unclear functional role. With many new structures to be generated by the structural genomics effort, new methods are needed to infer functional information from biomolecular shape, and numerous talks focused on novel methods of protein function prediction from structural data.

In some cases non-homologous proteins share functional elements that are very similar at the structural level. In these cases comparison of small motifs in protein structure provides a powerful method of function prediction. These predictions cannot be made from sequence analysis because they result from comparison of evolutionarily unrelated proteins. Brian Chen (Rice University, Houston, USA) described a new algorithm called 'match augmentation' for matching structural motifs, which is more efficient than currently available methods because it prioritizes the search by initially matching functionally significant residues. Chen and colleagues have also developed a strategy for estimating the statistical significance of structural matches and have shown that statistically significant similarities are functionally meaningful.

Purely geometric approaches for predicting various aspects of protein function were also described at the meeting. Two new computational geometry methods targeted the problem of protein-protein recognition. Yusu Wang (Duke University, Durham, USA) described a protein-docking algorithm based on the identification of protrusions and cavities on the surfaces of two proteins, which are aligned and scored with a simple scoring function. This algorithm for an initial rigid docking stage was able to generate near-native conformations for 24 out of 25 complexes from the Protein Data Bank. Xiang

Li (University of Illinois, Chicago, USA) presented a new empirical potential function for antigen-antibody recognition, developed with Jie Liang. The potential depends on local three-dimensional packing and is based on alpha-carbon shapes of antibody-antigen complexes. This potential was able to successfully recognize binding patches on the surfaces of native proteins. To facilitate the screening of phage-displayed combinatorial peptide libraries, Li and Liang have developed a method for designing biased peptide libraries enriched in native-like binding peptides.

Combining the evidence

We are now enjoying a wealth of highly diverse data at the genome-wide scale. Genomic sequences, protein structures, protein-interaction maps, gene-expression data, and data on protein-DNA binding all provide different perspectives on the molecular organization of the cell. Joint learning from these datasets will lead to new insights into the function of biological systems, and a variety of approaches to learning from these datasets were described, ranging from Bayesian networks to support vector machines to 'random forests'.

Tijl De Bie (Katholieke Universiteit Leuven, Leuven, Belgium) reported a method for predicting regulatory modules - that is, sets of transcriptional regulators together with their recognition sites and target genes. The method is the first to combine three independent sources of data: sequence motifs predicted by phylogenetic shadowing, chromatin immunoprecipitation followed by microarray analysis of the isolated DNA (ChIP-chip), and microarray gene-expression data. The method successfully predicted several known regulatory modules in yeast.

Several large experimentally and computationally derived datasets were similarly combined in a new method for predicting protein-protein interactions proposed by Yanjum Qi (Carnegie-Mellon University, Pittsburgh, USA). Many large datasets of protein-protein interactions in yeast are now available, but low coverage and very high false-positive rates are characteristic of most of the data on protein interactions. Qi and colleagues have shown that combining multiple sources of information improves the prediction of interacting protein pairs. To combine these diverse sources they adopt the so-called random forest technique, which uses a set of decision trees with random subsets of attributes. This method is used to compute similarity between protein pairs, and the k-nearest neighbor algorithm is then used to classify protein pairs as interacting or not. Tests showed that the method has 20% coverage at the 50% false-positive rate, which still compares favorably with previous approaches.

Understanding the individual genome

The vast amounts of information on DNA variation within populations have opened up new areas for the application of computational methods. Much of this variation is neutral in its effect on phenotype, and so it is essential to distinguish and understand that subset of genetic variation that does contribute to variation in phenotype. Phenotypically important single-nucleotide polymorphisms (SNPs) can be inferred from their predicted effect on molecular function and from the analysis of statistical signatures of natural selection in the genome. Computational approaches will potentially improve our understanding of the evolutionary mechanisms shaping genetic variation, will be useful for estimating the impact of polymorphic variants on gene function, and can be further applied in studies of the genetic basis of specific phenotypes.

Mutations are one source of DNA variation in the population, and an understanding of biochemical mechanisms of mutation is essential. Luciano Milanesi (Institute of Biomedical Technologies, CNR Milan, Italy) is part of an international collaboration looking for a link between chemical mechanisms of mutagenesis and the statistical properties of genetic variation. He described the analysis of several biochemical mechanisms leading to new mutations, which found that oxidative damage explains a large proportion of mutational hotspots. The analysis showed that the sequence context of a mutational hotspot is characteristic of a site of interaction with proteins involved in repair, replication or modification. Analysis of mutations induced by incorporation of the abnormal nucleotide 8-oxoGTP, which is produced by spontaneous oxidation of the guanine base in GTP *in vivo*, demonstrated that a substantial fraction of spontaneous AT to CT mutation is caused by 8-oxoGTP in the nucleotide pool.

Computational methods for predicting the phenotypic effect of amino-acid substitutions rely on various factors, including evolutionary conservation of the mutated position, accessible surface area of the mutated residue and other protein-structural parameters. Rachel Karchin (University of California, San Francisco, USA) described the use of mutual entropy to study structural and sequence features as predictors of the functional effect of sequence changes. She and colleagues employed a greedy algorithm, one that always follows a path that immediately increases the scoring function, to identify a subset of highly informative features from a set of 32 features. The usefulness of the selected features was demonstrated in a cross-validation test using a support vector machine. It was shown that a combination of solvent accessibility and evolutionary conservation gives as accurate a prediction of the functional effect of mutations as does the full set of 32 features.

Population genetic variation is one of the major factors responsible for differences in drug responses between individuals, and the emerging field of pharmacogenomics aims at developing personalized medicine adapted to an individual patient's genome. One of the challenges is to relate high-dimensional genomics data, such as microarray data on gene

expression, to clinical phenotypes. Jiang Gui (University of California, Davis, USA) described a method aimed at analyzing microarray data so as to select the expression of genes relevant to the survival of cancer patients. Based on a threshold gradient descent (TGD) method for the Cox regression analysis model, the method was applied to real data on survival after chemotherapy of patients with diffuse large B-cell lymphoma, and was shown to be useful for predicting survival and for identifying genes related to time to death.

Getting the name right

Many of the methods described at the meeting were attempts to predict functional features from genomic data. But what does one call these functional features and how does one describe the relationships between them? Without a well defined way to name aspects of biological function, genome annotation becomes a disorganized collection of chaotic irregular terms rather than a book of life. The development of a controlled vocabulary is essential for reasoning about biological data. Thus, it is not surprising that the topic of biomedical ontologies was included in the program for the third year in a row. Presentations described the creation of ontological resources and foundations of biomedical ontologies, integration of biomedical resources, and functional annotation.

Irena Spasic (University of Manchester, UK) presented a new measure for similarity between biological terms, which introduces an 'edit distance' to match the contexts associated with the terms. Edit distances will be familiar to bioinformaticians from the comparison of protein and DNA sequences, and are used here to identify similar terms in biomedical literature. The method showed good recognition of synonyms and is expected to facilitate the automated analysis of biomedical texts.

Merging existing terminology and ontology resources can result in new knowledge. Michael Cantor (Columbia University, New York, USA) is studying the relationship between diseases and genes. Using statistical and semantic relationships, he and colleagues have inferred relationships between disease concepts represented in the Unified Medical Language System (UMLS) and the Gene Ontology (GO). They used known gene-disease relationships from the Online Mendelian Inheritance in Man (OMIM) database to validate their approach, and they envisage that automated systems may eventually elucidate testable genetic hypothesis connecting clinical and biological knowledge.

Comparing this year's program with the programs of the first Pacific Symposia ten years ago, one can see that, although many new computational methods have emerged for analyzing new types of biological data, many traditional biologically motivated computational problems remain challenges for the field.